Supporting Information

Gianoulis *et al.* 10.1073/pnas.0808022106

SI Materials and Methods

GOS Data Collection and Preprocessing. For this study, we filtered the data from the first phase of the GOS expedition to keep only those sites that used a 0.1- to 0.8-μm filter size (with the exception of the Sargasso Sea station 11, which was excluded because it is suspected of contamination; see ref. 1); thus, only prokaryotes are part of this analysis. For the remaining 37 sites (Table S1), the site metadata were downloaded from the CAMERA database (2). For this study, measurements for temperature, sample depth, water depth, salinity, and monthly average chlorophyll level were used. Because 10 salinity measurements were missing, we averaged the salinity for all nonzero (excluded freshwater site) salinity measurements. In some cases, we were able to corroborate the missing measurements' validity through extrapolating from the World Ocean Database (3). For the protein sequence data, the 6.1 million predicted proteins (4) were downloaded from CAMERA.

Mapping Peptides to Sites. Peptides were mapped to sites based on the read-to-scaffold and orf-to-scaffold mappings available at CAMERA (2). Thus, to assign these peptides to a particular site, we used a mapping algorithm that cross-referenced between reads, scaffolds, and peptides based on predicted gene coordinates (Fig. S5). Therefore, there were instances in which reads that formed part of a single peptide originated from 2 different sites; because this allowed peptides to be "present" in multiple sites, we term these "multisite" peptides (for additional details, see below).

Mapping Cofactors for Modules. Cofactors were mapped to each module via EC number by using the Brenda database (5). To normalize the effects of module size, the fraction of chemical reactions requiring certain cofactors per module is regarded as the cofactor-dependence of module (Table S8). We then used a goodness of fit test (K-S test) to compare the distribution of canonical correlation analysis (CCA) structural correlation coefficient between the amino acids that have no cofactors (score = 0) and those with cofactors (score > 0) ($P < 0.05$).

Assignment and Pathway Score. The 111 Kyoto Encyclopedia of Genes and Genomes (KEGG) maps, 141 modules, and 191 operons were assigned as in ref. 6. For clarity, in the remainder of the text, we use the term pathway to refer to all of these levels. Module definitions were downloaded from KEGG (7), and operons were constructed as in ref. 8. In brief, protein sequences were searched against the extended database of proteins assigned to orthologous groups (OGs) in STRING 7.0 (8), by using BLASTP (9), and a pathway was called present when a hit matching 1 of its proteins occurred (with a BLAST score of at least 60 bits). All results described were also manually scrutinized to reduce artefactual assignments.

The pathway frequency for each site was assigned by summing the total number of instances of that pathway for a particular site and normalizing by total number of assignments for that site to compensate for sample coverage differences. For all analyses, pathways for which the summed count over all sites constituted equal to or <0.01% of the total count were removed to avoid artifacts.

In addition, we calculated a mismatch rate where we looked to see how many times the top 5 BLAST hits for each peptide mapped to the same pathway. We find that 80% of the top-5 hits will map to the same pathway with a corresponding drop at less stringent bit scores, suggesting our results are threshold-independent. A second source of miscalling could be cross hitting of pathways by more "generalist" enzymes. Therefore, we have manually checked the assignments and sought confirmation at multiple levels of resolution (map-module-operon-OG) for all of the case stories reported in this study.

Pairwise Correlations and Linear Regression. We computed pairwise Spearman correlations between each pathway frequency vector and each environmental metadata vector for the same sample set, (P values corrected for multiple testing by using the Benjamini–Hochberg false discovery rate; see ref. 10). Linear models were constructed in 2 directions: (*i*) the environmental factor was treated as the response variable and predicted from a subset of pathway frequencies; and (*ii*) the inverse model where pathway frequency was treated as the response variable and predicted from environmental factors. To identify the subset of predictive variables, we used a stepwise regression analysis based on Akaike's information criterion (implementation in R stats package). To avoid overfitting in (*i*), we used only the top 20 pathways that showed the highest pairwise correlation (as measured by uncorrected P value) with the environmental feature modeled. As in many feature selection methods, one is not guaranteed the "best" subset, and we acknowledge that there can be multiple suboptimal solutions. Linear models were considered significant at $P < 0.05$ for both the total model and the estimate of the variable coefficients. For regressions in both directions, the pathway frequencies were standardized to a mean of 0 and a SD of 1. For (*i*), we used the centered, quantile-normalized environmental data transformed into percentiles to ensure a truly normal distribution and, thus, accurate P values.

Clustering. The environmental data matrix was first standardized to mean of 0 and SD of 1. We evaluated distances by using 1-correlation and used average linkage hierarchical clustering. The clustering procedure was repeated by using spectral k-means without significance differences (data not shown).

Discriminative Partition Matching (DPM). To analyze whether groupings of sites based on similar environmental features also shared functional similarities, we clustered the sites based on their quantitative environmental metadata, resulting in 2 distinct clusters or site sets. Next, we partitioned the sites in the metabolism matrices (see Fig. 1*A*) into the same 2 site sets and calculated the mean normalized frequency for each pathway in each site set (see below for generalized approach). If the means of the pathway frequency between the 2 site sets were not significantly different, this would suggest that the environment-based partitioning does not reflect functional differences. If the distributions do differ significantly, it would imply that the environmental features are related to the specific aspect of metabolism. Also, we computed the 2-sample t test for each individual map, module, operons, and cluster of orthologous groups of proteins (COG). Those pathways that were significantly different (Benajamini–Hochberg corrected $P < 0.05$) were combined to form the DPM footprint (Table S4).

CCA. The goal of CCA is to identify the set of projections that maximally correlate 2 sets of variables (11). For a more detailed description of the relations of CCA to other common techniques, including principal components analysis and least squares regression, see ref. 12.

Due to the large number of dimensions and small number of data points, the solution can be unstable; thus, we applied a variant of CCA, regularized CCA (see ref. 13; implementation in ref. 14). We estimated regularization parameters λ_1 and λ_2 (penalty to covariance matrices) via a leave-one out cross-validation procedure (implementation in ref. 14; see Table S11). Because of the interdependencies between metabolic pathways, canonical weights must be interpreted with caution. For this reason, we also calculated the structural correlation coefficient, which is the correlation between the original variable and the canonical variate. This allows one to specifically answer the question how important is this one variable (metabolic pathway) relative to all of the other variables (metabolic pathways) (see below for additional evaluation metrics). Those pathways, which had a structural correlation coefficient >0.3, formed the CCA footprint (Table S5). Also, we investigated the effect of changing this threshold (see Table S11). Principal components analysis and the resultant biplot on the environmental features show these features to be basically orthogonal (Fig. S4).

Evaluation Metrics and Controls. ***Construction and results from control matrices.*** To control for relative differences in metabolic pathways among the geographic locations simply reflecting sampling bias, we constructed 2 control matrices, composed of proteins that would not be expected to change among sites, such as those involved in basal transcription or translational machinery. The first is composed of those COGs categorized as information processing, and the second, those involved in cellular processes.

We used Student's *t* test and found that, although the distributions of the means for the control matrices (composed of those COGs annotated as belonging to either information or cellular processing) are not significantly different between the 2 environmental site sets ($P = 0.07$ and 0.08, respectively), there are significant differences in metabolism ($P = 9 \times 10^{-3}$ and $4 \times 10^{-14,}$ COG and KEGG annotated metabolism definitions; see refs. 7 and 15). However, we do see the same asymmetry as originally noted in the GOS paper for DNA polymerase, topoisomerase, and gyrase (4), by aggregating across the basal machinery this effect is minimized. Thus, the greatest strength of DPM is as a means of evaluating the functional significance of a particular partitioning and in controlling for potential sampling bias through the testing of control matrices (expected to be environmentally invariant) alongside matrices that are suspected of being environmentally variant.

More detailed CCA evaluation metrics. As in PCA, there are a number of metrics that can be used to determine the number of dimensions, in this case canonical variates, that should be included in the analysis (12). The overall canonical correlations for both dimension 1 and dimension 2 are high for KEGG maps, module, and operons; however, there is a significant drop in average redundancy between dimension 1 and dimension 2 and further dimension 2 and dimension 3, making it appropriate to use only these 2 dimensions in the overall analysis (Table S11). We can also measure the amount of information the environment is able to "cover" from the environment and vice versa by calculating the average variance of the dimensions and the redundancy (11). These measurements are both high for the environment but lower for the metabolic pathways. This suggests that there are many weaker signals coming from the metabolic matrices as opposed to a few strong ones.

Generalization of DPM. We provided a specific use of DPM in the text; however, DPM can be generalized. There are 3 basic steps to DPM. (*i*) The sites from the first matrix are partitioned to create site sets. (*ii*) The second matrix is partitioned in accordance with these site sets. (*iii*) A *t* test (or ANOVA for more than 2 site sets) is performed to test whether the site sets are statistically different in the attributes of the second matrix.

Distributions of multisite and single-site peptides. In cases where a single peptide came from reads from different sites (multisite peptides) (16), we calculated the overlap between the reads and the peptide in 2 different ways. In the first, we considered the percentage of the read that was within a peptide (Eq. **1**), and in the second case, we assessed the amount that the read contributed to the peptide (Eq. **2**).

$$\text{Fraction of read within peptide} = \frac{\text{Read overlap with peptide}}{\text{Read length}} \quad [1]$$

$$\text{Fraction of peptide within read} = \frac{\text{Read overlap with peptide}}{\text{Peptide length}} \quad [2]$$

As illustrated in Fig. S3, the distributions of both peptide and read overlap are identical. This suggests that there are no major differences in assembly quality between the multisite and single site differences. However, it does not mean anything about the assembly quality itself. Only that if one assumes the assembly to be correct, there is no discernable differences between single and multisite peptides.

Comparison with Variance-Maximization Approaches. ***Compare and contrast with other methods.*** An entirely different approach to the one presented in the text assumes that the inherent variability of the environments could be directly observed by examining the global variance in the metabolic dataset; i.e., one identifies the pathways with greatest variance without directly measuring whether they covary specifically with the environment. First, we used standard deviation to find pathways that changed the most across the sites. We also used a PCA to identify the pathways that encapsulate the greatest proportion of variance. We then assessed the performance of these methods to identify metabolic adaptation to environmental parameters, based on their ability to recapture the environmental-based partitioning by using only the metabolic pathways identified as significant for each method by measuring cluster similarity (see below). Simply identifying the metabolic pathways with the greatest variance did not always reflect changes in the environmental parameters (see Fig. S2). Indeed, both methods that simultaneously incorporate environmental and metabolic data significantly outperform the variance-based, independent methods, and, perhaps, unsurprisingly, the linear models, which are more appropriate for investigating single relationships than looking at global context. These results were consistent, despite varying the number of pathways by using a variety of different thresholds for all methods SD, PCA, LM, DPM, and CCA (see Table S11).

Compare and contrast PCA and CCA. PCA and CCA actually have a deep relationship through formulation of the eigen problem, nicely illustrated in Borga *et al.* (12). Although they are related, there are completely different underlying assumptions motivating the use of one type and not the other. Although PCA attempts to capture the variance in a single dataset, CCA captures the within and between covariance (cross-variance) between 2 datasets. Thus, PCA can be used to extract components with the highest global variance, and has been used extensively in comparative metagenomics under the assumption that any variance observed could be attributed to environmental changes. Such reasoning makes sense when comparing qualitatively dissimilar environments. As an example, the difference between soil and water cannot really be quantified in a meaningful way, because all of the variables are changing simultaneously. Thus, more precise measurements to see how say metabolism varies as a function of the environment would not be

informative or even necessarily feasible. However, where environments change only slightly along a continuum of features, CCA provides the strength of covariation between both environmental and metabolic differences.

Metrics to compare cluster similarity. Cluster similarity was measured by computing both a normalized mutual information (NMI) score, which measures the amount of information lost if one applies the classification "clustering" from the first partition to the second (17), and the rand index (18), which computes the number of "correct" pairwise interactions between the 2 sets of clusters. The closer the NMI score is to 1.0 the better the metabolic footprint generated from the method performed in recapitulating the structure of the environmental data. For each set, a significance value for the rand index was computed by randomly shuffling the clustering assignment, recalculating the index after each iteration, and counting the number of times the index computed from the random data exceeded the index computed from the real data after 10,000 iterations.

1. Mahenthiralingam E, *et al.* (2006) Multilocus sequence typing breathes life into a microbial metagenome. *PLoS ONE* 1:e17.
2. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: A community resource for metagenomics. *PLoS Biol* 5:e75.
3. Boyer T, *et al.* (2006) World Ocean Database 2005, in *NOAA Atlas NESDIS 60* (U.S. Government Printing Office, Washington, DC), p 190.
4. Yooseph S, *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PLoS Biol* 5:e16.
5. Barthelmes J, Ebeling C, Chang A, Schomburg I, Schomburg D (2007) BRENDA, AMENDA and FRENDA: The enzyme information system in 2007. *Nucleic Acids Res* 35:D511–514.
6. Tringe SG, *et al.* (2005) Comparative metagenomics of microbial communities. *Science* 308:554–557.
7. Kanehisa M, *et al.* (2006) From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res* 34:D354–357.
8. von Mering C, *et al.* (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* 35:D358–D362.
9. Altschul SF, *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
10. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Stat Soc B* 57:289–300.
11. Wichern R, Johnson D (2003) *Applied Multivariate Statistical Analysis* (Prentice Hall, Upper Saddle River, NJ), 5th Ed.
12. Borga M (1998) *Learning Multidimensional Signal Processing. PhD* (Linkoping University, Linkoping, Sweden).
13. Eaton M, Perlman M (1973) The Non-Singularity of Generalized Sample Covariance Matrices. *Ann Stat* 1:710–717.
14. Gonzalez I, Dejean S, Martin PG, Baccini A (2008) CCA: A R package to extend canonical correlation analysis. *J Stat Softw* 22:1–14.
15. Tatusov RL, *et al.* (2003) The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4:41.
16. Rusch DB, *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5:e77.
17. Forbes A (1995) Classification-algorithm evaluation: Five performance measures based on confusion matrices. *J Clin Monitor Comp* 11:189–206.
18. Hotelling H (1936) Relations between two sets of variates. *Biometrika* 28:321–377.

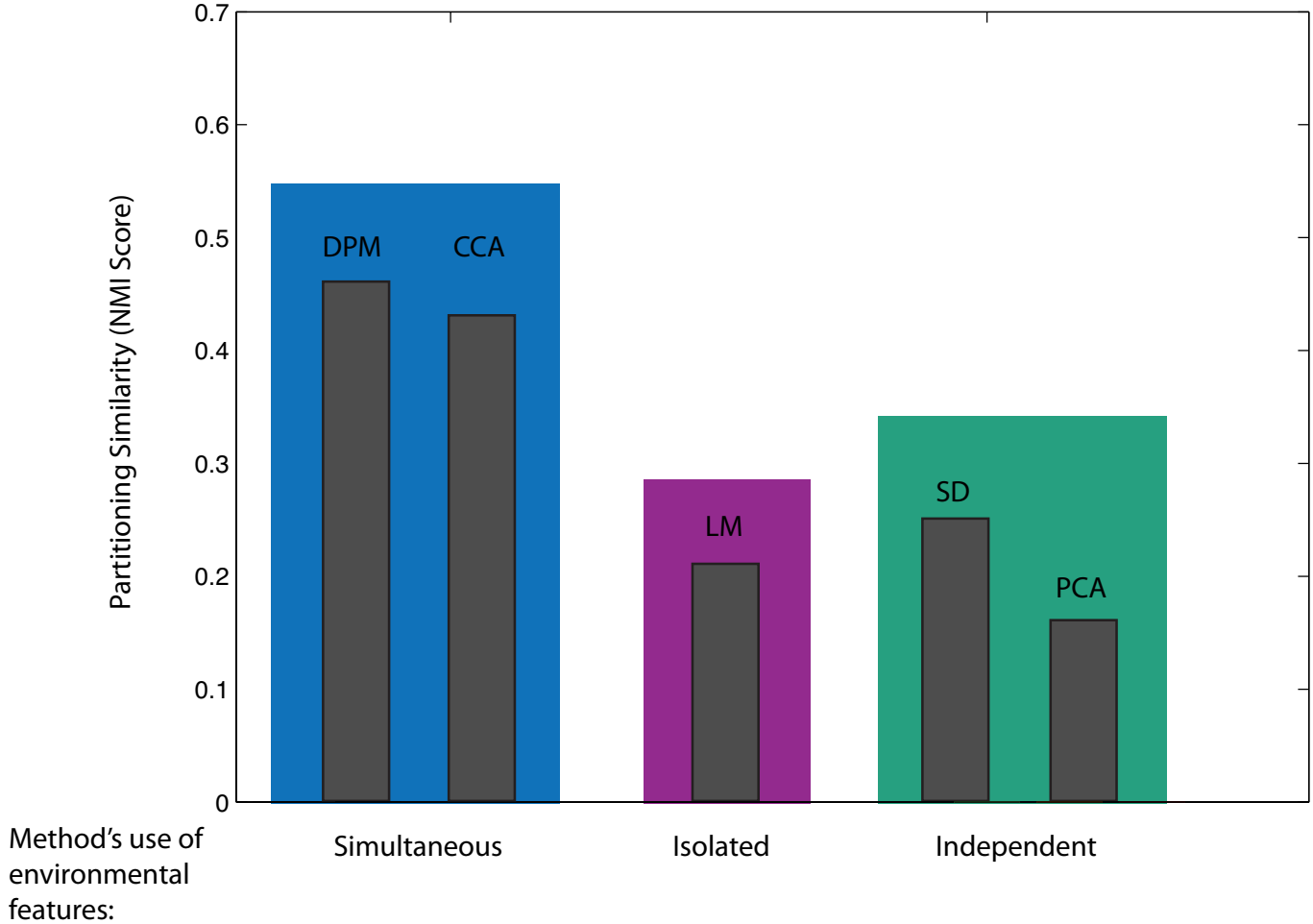


Fig. S1. Comparison of different classes of methods. We evaluated the efficacy of 3 different classes of methods based on their explicit use of the quantitative environmental data, which we term independent, isolated, and simultaneous. Independent methods include no environmental description (green), isolated only one environmental feature at a time (purple), and simultaneous methods incorporate all environmental features simultaneously (blue). For clarity, we refer to the highly-weighted set of pathways generated for each method as a footprint. Each of the 5 methods was used to generate a metabolic footprint, and each bar represents the NMI score for that footprint of method. No statistically significant difference was observed between scores within each particular category ($P > 0.05$).

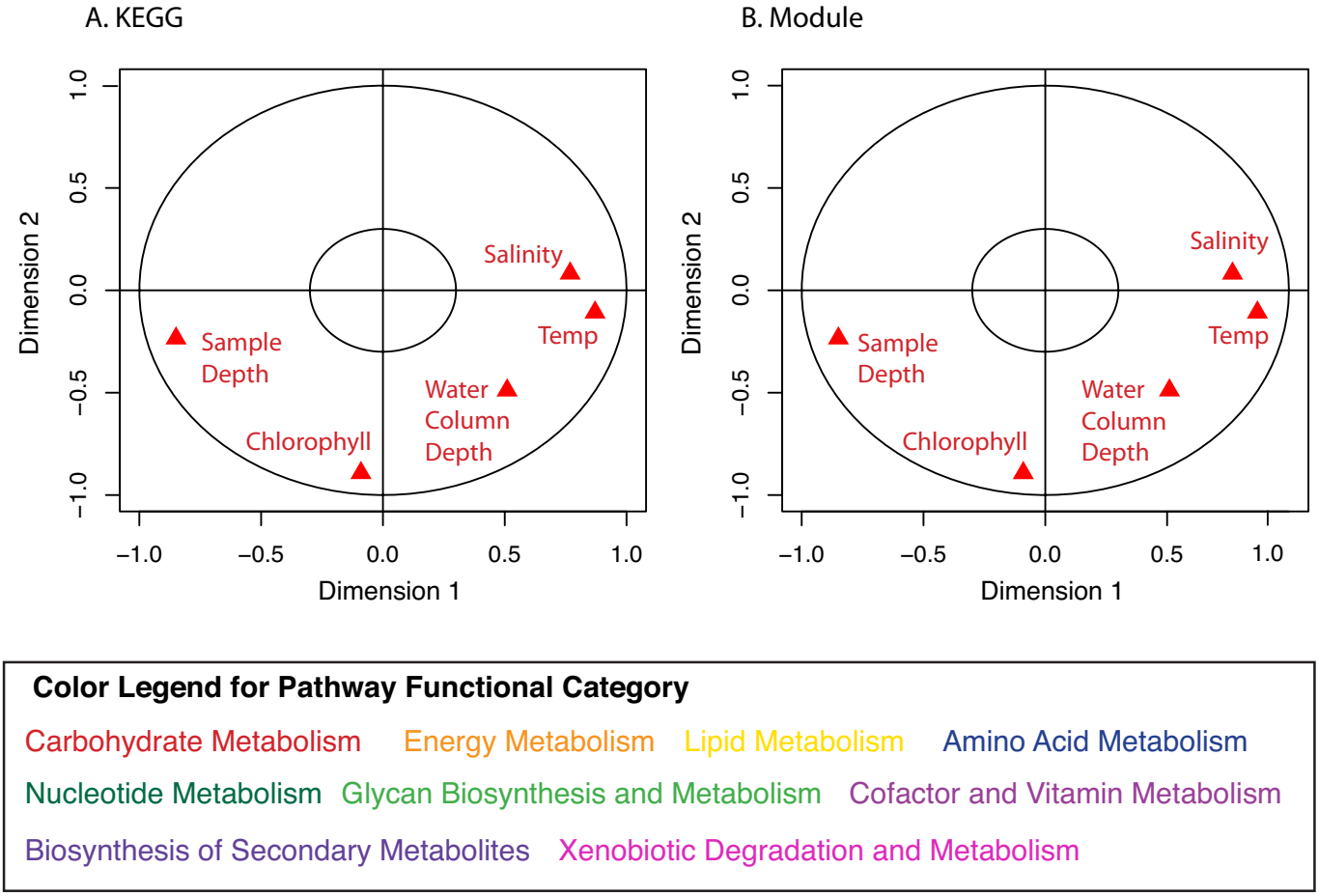


Fig. S2. Bullseye plot of CCA-derived structural correlations. Shown are results from CCA for KEGG (*A*) and module (*B*). The *x* and *y* axes represent the structural correlation coefficients (normalized weights) in the first and second dimension, respectively. The closer either environmental features (red triangles) or metabolic pathways (color coded by functional category) are to the perimeter of the outer ellipse, the better they fit the model. Also, the closer an environmental feature is to a metabolic pathway the stronger the covariation between them. The inner ellipse (radius 0.3) represents those features that did not fit the model (for further explanation, see ref. 14). Those pathways in the inner ellipse can be thought of as environmentally invariant, and those outside this ellipse as environmentally variant.

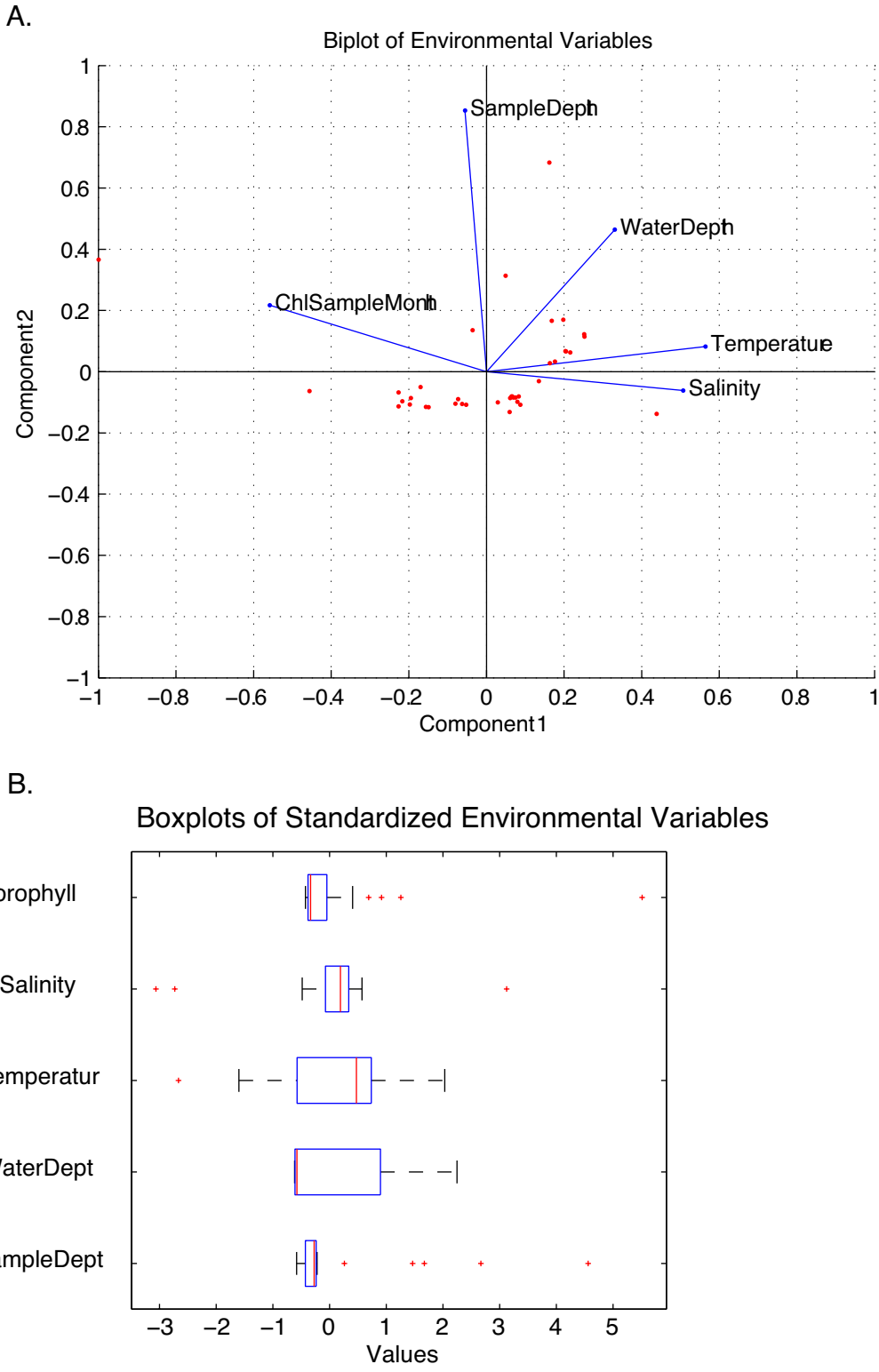


Fig. S3. Biplot and boxplot of standardized environmental variables. To examine possible dependencies between the variables, we performed principal-component analysis. (*A*) We next plotted component 1 and 2. One can see that the variables with the exception of temperature and salinity are basically orthogonal to one another. (*B*) Boxplot of standardized variables.

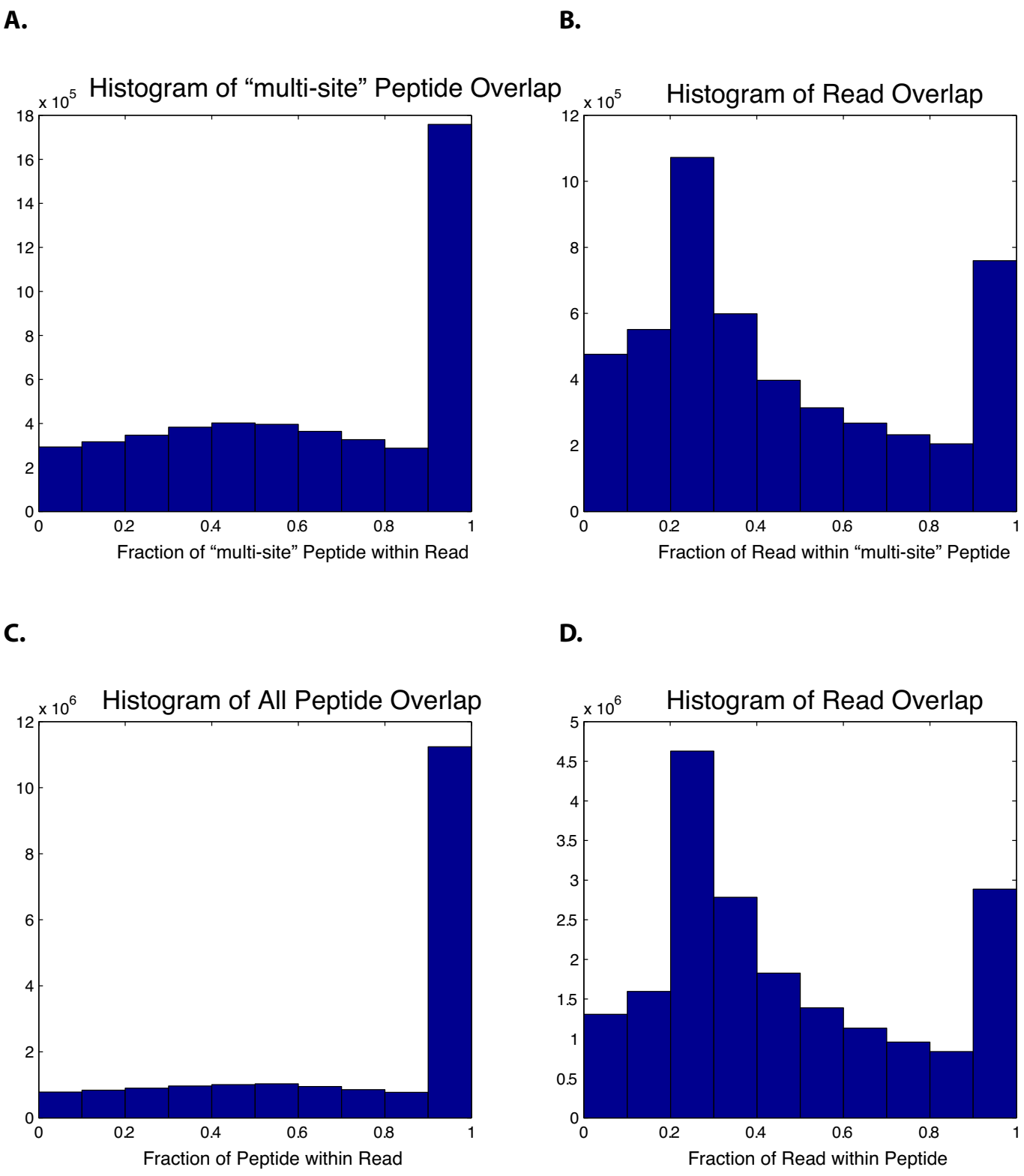
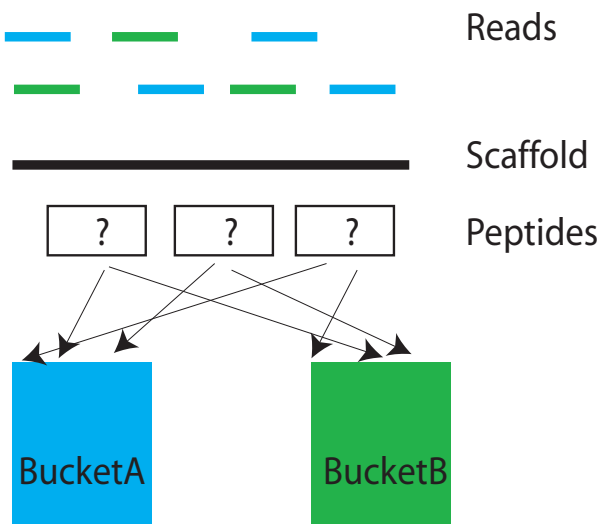


Fig. S4. Distribution of multisite and all-peptide and read overlaps. We term peptides whose reads derived from multiple geographic locations ''multisites.'' (*A* and *B*) Fraction of each peptide that is within read (*A*) and the fraction the read ''contributes'' to the peptide (*B*). The same plots are shown for all peptides: peptide overlap (*C*) and read overlap (*D*). While *B* and *D* are in part assessing if differences in assembly between multi- and single-site peptides exist, *A* and *C* illustrate more functionally that there is no observable difference between 2 types.



Given a scaffold where
R - set of reads
BS- set of buckets to which reads from R belong
P - set of peptides

Pseudocode:

```
foreach p in P:
        s=p's scaffold
        find R(s)
        foreach r in R(s):
                if r overlaps with p
                        put p in r's bucket
```

Fig. S5. Mapping peptides to geographic locations (sites). Schematic and pseudocode for mapping of peptides to a particular site are shown. The goal of this algorithm is, given this set of reads [color coded blue or green, depending on which bucket (site) they were recovered from], this set of peptides (boxes), and the coordinates from the scaffold (long black line), to determine to which buckets the peptides (boxes) belong.

Other Supporting Information Files

Table S1
Table S2
Table S3
Table S4
Table S5
Table S6
Table S7
Table S8
Table S9
Table S10
Table S11